04-19-00

A

THE COMMISSIONER OF PATENTS AND TRADEMARKS
Washington, D.C. 20231

Case Docket 940630-010-020
Date: April 18, 2000

Sir:

Transmitted herewith for filing is the Patent Application of:

Inventor(s): Roma, Norbert

For: A METHOD AND APPARATUS FOR COMPARING SCORES IN A VECTOR SPACE RETRIEVAL PROCESS

Enclosed are:

[X] 2 sheets of informal drawing(s).
[ ] A certified copy of a _____ application.
[X] Declaration and Power of Attorney.

[ ] PTO Form 1449 with attached references and transmittal letter.

The filing fee has been calculated as shown below:

For:	No. Filed	No. Extra	Rate	Fee
Basic Fee				$690.00
Total Claims	20 - 20 =	x 0	x 18 =	$ 0
Indep Claims	5 - 3 =	x 2	x 78 =	$156.00
[ ] Multiple Dependent Claim Presented			+ 230 =	$ 0
			TOTAL	$846.00

[X] A check in the amount of $846.00 to cover the filing fee is enclosed.
[ ] Please charge my Deposit Account No. 10-1202 in the amount of $ _____. A duplicate copy of this sheet is enclosed.
[X] The Commissioner is hereby authorized to charge payment of the following fees associated with this communication or credit any overpayment to Deposit Account No. 10-1202. A duplicate copy of this sheet is enclosed.
[X] Any additional filing fees required under 37 CFR 1.16.
[X] Any patent application processing fees under 37 CFR 1.17.

Respectfully Submitted,

By _____
Attorney: Blaney Harper
Registration No.: 33,697
Telephone: (202) 879-7623

Jones Day Reavis and Pogue
51 Louisiana Avenue, N.W.
Washington, DC 20001-2113

Express Label No.
EJ844955269US

Atty. Docket No. 940630-010-020

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of: Roma, Norbert

Serial No.: Not Assigned Group Art Unit: Not Assigned

Filed: Herewith Examiner: Not Assigned

For: A METHOD AND APPARATUS FOR COMPARING SCORES IN A VECTOR SPACE RETRIEVAL PROCESS

Commissioner of Patents and Trademarks
Washington, D.C. 20231

EXPRESS MAIL CERTIFICATE

"Express Mail" label number: EJ844955269US

Date of Deposit: April 18, 2000

I hereby certify that the following **attached** paper or fee

- Patent Application Transmittal Letter
- Patent Application Specification
- Combined Declaration and Power of Attorney
- Patent Application Fee

is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Commissioner of Patents and Trademarks, Box Patent Application, Washington, D.C. 20231.

Blaney Harper
(Name of person mailing paper or fee)

(Signature of person mailing paper or fee)

Jones, Day, Reavis & Pogue
51 Louisiana Avenue, N.W.
Washington, DC 20001-2113

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

UTILITY
PATENT APPLICATION
TRANSMITTAL

Atty. Docket No. 940630-010-020

First Inventor or Application Identifier

Roma, Norbert

Title:

A METHOD AND APPARATUS FOR COMPARING SCORES IN A VECTOR SPACE RETRIEVAL PROCESS

Express Mail Label No.: EJ844955269US

Assistant Commissioner for Patents
Box Patent Application
Washington, D.C. 20231

FILING UNDER 37 CFR §1.53(b)

1. ☒ *Fee Transmittal Form (e.g.,PTO/SB/17)
2. ☒ Specification [Total Pages 18]

   (preferred arrangement set forth below)

   -Descriptive title of the Invention
   -Cross References to Related Applications
   -Background of the Invention
   -Brief Summary of the Invention
   -Brief Description of the Drawings
   -Detailed Description
   -Claim(s)
   -Abstract of the Disclosure
3. ☒ Drawing(s) (35 U.S.C. § 113) [Total Sheets 2]
4. Oath or Declaration [Total Pages 3]
   a. ☐ Newly executed (original or copy)
   b. ☐ Copy from a prior application.

15. ☐ . Other:

16. **If a Continuing Application,** check appropriate box, and supply the requisite information below and in a preliminary amendment:

☐ Continuation ☐ Divisional ☐ Continuation-in-part (CIP)

of prior application No.: _____

Prior application information: Examiner _____ Group/Art Unit: _____

For CONTINUATION or DIVISIONAL APPS only: The entire disclosure of the prior application, from which an oath or declaration is supplied under Box 4b, is considered a part of the disclosure of the accompanying continuation or divisional application and is hereby incorporated by reference. The incorporation can only be relied upon when a portion has been inadvertently omitted from the submitted application parts.

17. CORRESPONDENCE ADDRESS

☐ Customer Number or Bar Code Label

or ☒ Correspondence address below

Blaney Harper
Jones, Day, Reavis and Pogue
51 Louisiana, N.W.
Washington, DC 20001-2113
Telephone: 202/879-3939
Facsimile: 202/626-1700

I hereby declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issues thereon.

April 18, 2000
Date

By Blaney Harper
Blaney Harper - Attorney
Reg. No. 33,897

Jones, Day, Reavis and Pogue
51 Louisiana Ave., NW
Washington, DC 20001-2113
(202) 879-3939

Title of The Invention

A Method And Apparatus For Comparing Scores In A Vector Space Retrieval Process

Field of The Invention

The present invention relates to the field of computerized search and retrieval systems. More particularly, this invention relates to a method and apparatus for retrieving information wherein a vector space search algorithm is used to retrieve information concerning multiple vector profiles.

Background of The Invention

Advances in electronic storage technology has resulted in the creation of vast databases of documents stored in electronic form. These databases can be accessed from remote locations around the world. As a result, vast amounts of information are available to a wide variety of individuals. Moreover, information is not only stored in electronic form, but it is created in electronic form and disseminated throughout the world. Sources for the electronic creation of such information includes news, periodicals, as well as radio, television and Internet services. All of this information is also made available to the world through computer networks, such as the worldwide web, on a real time basis. The problem with this proliferation of electronic information, however, is how any one individual may access information useful to that individual in a timely manner. In particular, how any one individual can receive individual pieces of information on a real time basis (*e.g.*, a stream of documents) and decide which pieces of information are useful to the user.

Specifically, there are many search techniques to retrieve information from a database or data stream such as Boolean word searches, typed information retrieval or vector space based retrieval algorithms. Vector space based algorithms calculate a number that represents the

similarity between any document in a database and a vector profile having a series of terms or phrases. Vector space based algorithms, while general and sophisticated, have several shortcomings. One of them is the fact that numeric vector space scores of documents against two different profiles, in general, are not directly comparable to each other. This is unsatisfactory for several reasons. First, from the point of view of an end-user, it might be desirable to inspect scores for a certain document in contexts of several profiles. This could be done, for example, in order to evaluate the performance of the profiles in question so that they can be adjusted to improve their accuracy. Another use for comparable, or normalized, scores across profiles is to facilitate a multiple classification procedure. One way to implement a multiple classifier is by employing a score threshold for tags (classes, profiles). For this to be meaningful, the scores for different tags have to be comparable to each other.

Objects of The Invention

It is an object of the present invention to provide an improved method and apparatus for retrieving information from a data source such as a data stream.

It is another object of the present invention to retrieve information by comparing scores of multiple vector profiles.

It is still a further object of the present invention to make use of the score distribution of different profiles to compare scores.

Brief Description of The Drawings

Figure 1 is a block diagram that illustrates a computer system for performing information extraction according to one embodiment of the present invention.

Figure 2 illustrates the scaling of delivery ratio maps according to one embodiment of the present invention.

Figure 3 illustrates the scaling of delivery ratio maps according to another embodiment of the present invention.

Summary of The Invention

In general, normalizing scores relies on the existence of a natural normalization reference in form of a reference collection. For example, an organization interested in classifying news would be able to collect an archive of, for example, a week's worth of news articles. This is a natural resource with general properties of the news stream that can be used to normalize the scores of profiles classifying this stream. One possible way to draw on this resource is to use a delivery ratio mechanism to establish a set of, for example, nine (vector space) score thresholds (let's call them $x_k, k = 1, 2, ..., 9$) for each profile. Now, each new article in the stream can be assigned to one of the ten levels delimited by those thresholds. These levels can be treated as a discrete, ten-point score for the document. Clearly the score is comparable across profiles.

The delivery ratio of r (which is a fraction between 0 and 1) partitions a stream (or a static set - these two kinds of document sources are interchangeable) of documents into a section of top scoring r-fraction of documents and the remainder. This way a set of successively bigger delivery ratios, $r_1, r_2, r_3, ...$ sections the stream into tiers. Any given document is assigned to a tier according to how many delivery ratio thresholds it matched or surpassed and how many it failed to reach. This creates a scoring structure which reflects the specificity of the document with respect to a profile in terms of density of relevant documents in the stream. In other words, a document in the k^{th} tier is such that it failed to be classified in the top r_k ratio of the stream (thus r_k fraction of the stream is more relevant to the given profile than the document under consideration). At the same time this document was classified as being in the top r_{k-1} part of the

stream. Thus, this mechanism defines a score (referred to as σ) for a document depending on how it compares to other documents in the stream when scored against a given profile.

Description of The Preferred Embodiments

Figure 1 is a block diagram of a computer system used for retrieving information from a database. Computer 20 comprises a central processing unit (CPU) 30 and main memory 40. Computer 20 is connected to an Input/Output (I/O) system 10 and disk storage unit 50. The I/O system 10 includes a display 5, a keyboard 7 and a mouse 9. Furthermore, this computer system is connected to a variety of networks for communicating with other computers and obtaining access to remote databases. Among the networks connected to this computer system is the worldwide web 55, an intranet 57, private external network 59. In general, the disk storage unit 50 stores the program for operating the computer system and it stores the documents of the database. The computer 20 interacts with the I/O system 10 and the disk storage unit 50. The computer 20 executes operations according to instructions contained in a program that is retrieved from the disk storage unit 50. This program causes the computer 20 to retrieve the text of documents, or parts thereof, which are stored in a database located either in disk storage 50 or in a storage location accessible over a network. These instructions also cause information received over a network to be distributed to specific individuals over a network based on the content of the information.

According to the present invention, the program that performs the information extraction incorporates instructions that compare scores from a source document against scores obtained from reference data. To consider the problem of score comparison, the characteristics of the source data - either a static corpus (but only as a stream-like source of documents, in other words we do not assume the ability to do retrieval on the document source), or a live document stream

are known. Additionally, a reference corpus with term-statistics compatible with that of the document source exists. To define the normalized score, the following series of steps is performed.

1. The first parameter necessary to build the score is the number of thresholds used, denoted n. This count of thresholds includes all the (non-trivial) thresholds $r \in (0,1)$. Thus, if $r_0 = 0$, the set of all thresholds in the score is indexed: r_k where $k = 0, 1, ..., n$. In addition, imposing $r_{n+1} = 1$ ensures that the highest tier's size is consistent with others.

2. Choose a mapping $r: \{1, 2, ..., n\} \to (0, 1)$ from the set of indices, as discussed below, to set all possible delivery-ratio thresholds.

3. In the next step we need to run a set of trained profiles against the reference corpus. This run produces (via the delivery ratio mechanism used for each r_k for each profile) a set of vector space scores, $x_k, k = 1, 2, ..., n$ corresponding to the thresholds r_k for each profile in turn.

4. Now, we are ready to score documents from the document source: Given an incoming document, we obtain a vector space score, x, of that document against a given profile, and we compare it to all the thresholds x_k. Let $l = 1, 2, ..., n - 1$ be such that $x_l \le x < x_{l+1}$, then we assign a score $\sigma = l$ to that document. For documents with $x < x_1$ we set the score $\sigma = 0$, and in case of $x \ge x_n$ the score is $\sigma = n$. This is the normalized score of the document against the given profile. Of course, the actual numerical value of σ can be defined in any number of ways, but the choice presented here is simple, and easy to interpret.

Furthermore, certain mappings are defined. The mapping used to set delivery ratio thresholds is defined as follows:

$$r_k = \frac{1 - a^{-k}}{1 - a^{-(n+1)}}$$

and is parameterized by the base $a \in (1, \infty)$, of the exponent. This parameterization allows for relative scaling of maps. Figure 2 demonstrates the scaling of maps corresponding to $a = 1.5$ (100) and $a = 4$ (150) where n = 10.

The power law mapping also uses one free parameter, s (which is used for relative scaling) and is defined as follows:

$$r_k = \left(\frac{k}{n+1} \right)^{\frac{1}{s}},$$

where $s \in (1, \infty)$. This formula normally gives practical mappings for s>4, and is characterized by the fact that it is close to linear at the top of the scale (for the high-scoring documents). Figure 3 illustrates the power law character of the curve on two examples wherein S = 4 (200) and S = 7 (250). The nearly linear behavior of the scale at the high end of the scores may not be the best way to space the thresholds from the point of view of direct user judgment, but may be very useful in automatic systems.

It is clear from the definitions of the maps above that one could extend the normalized score from a discrete set of values to a continuous scale. This is accomplished in several ways:

1. The exponential scale above can be redefined, using another integer parameter $m>n$, such that now $k = 0, 1, ..., m$ and

$$r_k = \frac{1 - a^{\frac{-k(n+1)}{m}}}{1 - a^{-(n+1)}}$$

This gives us the old mapping for $m = n + 1$, but gives us a continuous scale in the limit of $m \to \infty$, so it can be approximated through using large m.

2. In case of the power law, simply taking $n \to \infty$ produces a continuous scale.

3. The above two methods call for calculation of a large number of thresholds for each scale, which is computationally undesirable. It may be more practical to use a discrete scale and linearly interpolate between the discrete scores to obtain an approximate continuous scale. Specifically, if we choose to associate a score $\sigma(k) = k$ (as defined above,) with each x_k (for $k = 0, 1, ..., n$), then we can define the interpolated $\sigma(x)$ for $x_k < x < x_{k+1}$ as follows:

$$\sigma(x) = \frac{x - x_k}{x_{k+1} - x_k}(\sigma(k+1) - \sigma(k)) + \sigma(k).$$

This defines a continuous scale of scores σ between $\sigma = 0$ corresponding to $x_0 = 0$, and $\sigma_{n+1} = n + 1$ corresponding to x_{n+1}. The definition of x_{n+1} is the highest possible score of a document against the given profile. This is a well defined number for most vector space scoring algorithms.

Finally, there is the subject of selection of the scaling parameters present in the mappings introduced above. (a and s for exponential and power law maps respectively.) The easiest case is when we have relevance judgments available for a sizable *random* sample of the reference

collection. In this scenario we can simply adjust the specificity of a set of profiles by looking at the total numbers of documents relevant to each profile in the sample. The ratios of those totals would determine the appropriate ratios of scaling parameters for those profiles. (Specifically, the ratio of totals could be set equal to a ratio of two r_k's for two different profiles at some fixed k, corresponding to some fixed $\sigma = k$.)

This new normalization method may be used for filtering applications, and it can be easily adapted to scoring interest representations against any document source. For example, a company may have a set of (say, 100) different tags that it would like to assign to articles in the incoming news stream. It is reasonable to assume that such a company would be able to collect an archive of, say, a month's worth of news. Furthermore, it is reasonable to expect that the general characteristics of the news stream change slowly enough, that the sample would continue to be a valid representation of the term usage in the news stream for many months.

With the corpus of archived articles ready, it is possible now to create a score scale for each profile according to the method described above. This allows for immediate application of the delivery ratio method for thresholding, as well as any other thresholding mechanism that relies on the newly produced scores. Moreover, the scores for any given news item are comparable across all profiles, which makes it possible to set threshold performance for tags. Thus a variable number of (rank-ordered) tags can be assigned to any given news story, depending on how many profiles score above the threshold. This technique of assigning a well controlled number of multiple tags to a document is usually called multiple classification.

This threshold, of course, may be tag-dependent, which, for example in the case of the delivery ratio thresholding, could be a reflection of different specificity of various tags. To illustrate this let me consider three example tags: "sports", "ice hockey", and "NASA". In a

general news stream, we could reasonably expect that around 10% of the news stories would have to do with sports to a degree that would warrant assigning the tag "sports" to them. On the other hand, it is equally reasonable to think that no more than, say, 1% of the news stories were sufficiently relevant to the subject of ice hockey to justify assigning the corresponding tag. This tighter threshold for "ice hockey" reflects the fact that it is a subset of sports, and so it is more specific (or focused). Of course, two profiles do not need to include one another to differ in the level of specificity: As before, no more than 1% of the stories in the general news stream are likely to need the tag "NASA", but this time the reason for the low number is not that the subject is a subset of "sports", but solely because the definition of the interest is more limited.

While assigning individual thresholds to tags does help solve the issue of tag specificity, a more flexible way of addressing this problem can be found through scaling the whole score scale for a tag. For example, instead of using different thresholds for "sports" and "ice hockey" tags, we could use one score threshold, θ, and calibrate the scores for both tags so that θ corresponded to a 10% point for "sports" and to 1% for "ice hockey".

First, the number of score points must be determined. A scale with ten distinct scores, referred to as $\sigma = 0, 1, 2, ..., 9$ is used. As a result, setting θ to the σ score of 3, provides flexibility so that there are three score levels among the rejected articles (those scoring 0 - completely irrelevant, 1 - remotely related, or 2 - "near misses"), and a (partially) ranked list of relevant news items. For the tag "sports", this means that the third (since $\theta = 3$) threshold in the score scale $r_3 = 0.9$, because the top 10% of the news stream are treated as relevant to the tag. Based on that requirement, the base of the exponent is determined and then the other eight thresholds in the score are computed. Carrying out the computations produces the following sequence of delivery ratio thresholds, each of which corresponds to respective σ scores:

σ	0	0.5354	0.7843	0.9000	0.9538	0.9790	0.9904	0.9958	0.9983	0.9995
r_k	0	1	2	3	4	5	6	7	8	9

Now, we can run the profile for "sports" in retrieval mode against the reference corpus and, using the delivery ratio algorithm, find the raw vector space scores, x_k, corresponding to all the r_k thresholds. This way, once we start scoring incoming news stories, we will be able to assign a σ score to each one of them based on its vector space score, and the number of thresholds in the normalized scale that it passed. This is the complete recipe for generating normalized score for news stories about sports events. To complete the example with an illustration of relative scaling, let's go back to the "ice hockey" tag.

We can treat the tag "ice hockey" much the way we dealt with "sports", with the difference being that now we impose $r_3 = 0.99$, since only 1% of the news stream is expected to be relevant to the topic. This, again, results in a sequence of delivery ratio thresholds, which are then used to assign the σ score to each news article. (As before, we need to run the profile in retrieval mode against the reference database to obtain the vector space scores, x'_k, for each of the thresholds r_k' for "ice hockey".) The following table demonstrates how a uniform score, σ, is associated with different sets of thresholds for the two profiles we are considering:

σ	sports r_k	ice hockey r'_k
0	0	0
1	0.5354	0.784555
2	0.7843	0.953584
3	0.9000	0.990000
4	0.9538	0.997846
5	0.9790	0.999536

6	0.9904	0.999900
7	0.9958	0.999979
8	0.9983	0.999996
9	0.9995	0.999999

The above example demonstrates the following steps in the procedure for deriving a normalized score for a collection of profiles, using a reference database:

1. Choose the number of uniform score levels to appropriately reflect the distribution of relevance over the source stream. (Set n.)
2. Decide on which mapping (for example, chose one of the two such mappings discussed above) is more appropriate for the given application.
3. Select the threshold of relevance on the new scale for each tag. (Set $\theta = 3$. Optionally impose $r_3 = 0.9$; Instead of setting one of the r_k 's to a given ratio, one could just fix the base of the exponent, which would in turn determine r_3.)
4. Run the profile for the tag in retrieval mode against the reference collection in order to get the raw vector space scores for the delivery ratio thresholds comprising the score scale.
5. For each new story, we can now determine its vector space score, compare it to the delivery ratio scores in the normalized score scale, and, depending on how many of those thresholds it passes, we can assign it a normalized (σ) score, and compare it to θ. This last comparison determines whether the news story receives the tag or not.
6. As an extra bonus, we end up with a rank-ordered list of tags. This can be used to limit the number of tags to a certain maximum (even if more tags manage to pass the threshold of θ) of most relevant tags. It can also be used to assign several highest scoring tags to documents that did not score above θ for any tags, as long as some tags scored above 0.

The above example demonstrates in concrete terms how to implement the normalized score method in the tagging application scenario. While this invention has been particularly described with reference to specific embodiments, those of skill in the art will recognize that changes to specific embodiments may be made without departing from the spirit and scope of the present invention.

CLAIMS

I claim:

1. A method of selecting documents from a data stream, comprising:

selecting a resource having information comparable to said data stream:

selecting at least one topic;

analyzing said topic against said resource;

analyzing said topic against said data stream; and

comparing results from said data stream analysis to results from said resource analysis to select a document from said data stream.

2. A method of selecting documents from a data stream, comprising:

selecting a profile;

analyzing a reference corpus of documents against said profile to determine at least one score;

scoring at least one document from said data stream against said profile; and

comparing said scores from said data stream document to said at least one score from said reference corpus to select said document from said data stream.

3. A method as in claim 2, further comprising:

determining a plurality of reference corpus scores defining a plurality of delivery ratios; and

determining a delivery ratio that corresponds to said score from said data stream document to select said data stream document.

4. A method as in claim 3, wherein said delivery ratios correspond to said reference corpus scores according to an exponential decay function.

5. A method as in claim 4, wherein said exponential decay function is defined as:

$$r_k = \frac{1 - a^{-k}}{1 - a^{-(n+1)}}$$

wherein k, correspond to an integer, $a \in (1, \infty)$, and r_k corresponds to a delivery ratio.

6. A method as in claim 3, wherein said delivery ratios correspond to said reference corpus scores according to a power law function.

7. A method, as in claim 6, wherein said power law function is defined as:

$r_k = (K/(N+1))^{(1/S)}$, wherein $S \in (1, \infty)$.

8. A method of retrieving information from a data source, comprising:

receiving an information request from a communications network;

selecting a data source;

selecting a resource having information comparable to said selected data source;

selecting at least one topic;

analyzing said topic against said resource;

analyzing said topic against said selected data source; and

comparing results from said selected data source analysis to results from said resource analysis to retrieve at least one document from said selected data source; and

transmitting said retrieved documents over said communications network.

9. A method of retrieving information from a data source, comprising:

receiving an information request from a communications network;

selecting a data source;

selecting a profile;

analyzing a reference corpus of documents against said profile to determine at least one score;

scoring at least one document from said selected data source against said profile; and

comparing said scores from said selected data source documents to said at least one score from said reference corpus to retrieve at least one document from said selected data source; and

transmitting said retrieved documents over said communications network.

10. A method as in claim 9, further comprising:

determining a plurality of reference corpus scores defining a plurality of delivery ratios; and

determining a delivery ratio that corresponds to said score from said data stream document to select said data stream document.

11. A method as in claim 10, wherein said delivery ratios correspond to said reference corpus scores according to an exponential decay function.

12. A method as in claim 11, wherein said exponential decay function is defined as:

$$r_k = \frac{1 - a^{-k}}{1 - a^{-(n+1)}}$$

wherein k, correspond to an integer, $a \in (1, \infty)$, and r_k corresponds to a delivery ratio.

13. A method as in claim 10, wherein said delivery ratios correspond to said reference corpus scores according to a power law function.

14. A method, as in claim 13, wherein said power law function is defined as:

$r_k = (K/(N+1))^{(1/S)}$, wherein $S \in (1, \infty)$.

15. A computer system for retrieving information from a data source, comprising:

a central processing unit coupled to a memory unit, an input system and a communications network;

said central processing unit executes instructions retrieved from said memory in response to commands entered into said input system, said central processing unit transmits a request over said communications network, said request causes a computer system receiving said request to:

i) select a data source;

ii) select a profile;

iii) analyze a reference corpus of documents against said profile to determine at least one score;

iv) score at least one document from said selected data source against said profile;

v) compare said scores from said selected data source documents to said at least one score from said reference corpus to select at least one document from said selected data source; and

vi) transmit said selected documents over said communications network; and

said central processing unit executes instructions to retrieve said selected documents from said communications network.

16. A system, as in claim 15, wherein said receiving computer system:

determines a plurality of reference corpus scores defining a plurality of delivery ratios; and

determines a delivery ratio that corresponds to said score from said data stream document to select said data stream document.

17. A system as in claim 16, wherein said delivery ratios correspond to said reference corpus scores according to an exponential decay function.

18. A method as in claim 17, wherein said exponential decay function is defined as:

$$r_k = \frac{1 - a^{-k}}{1 - a^{-(n+1)}}$$

wherein k, correspond to an integer, $a \in (1, \infty)$, and r_k corresponds to a delivery ratio.

19. A method as in claim 17, wherein said delivery ratios correspond to said reference corpus scores according to a power law function.

20. A method, as in claim 19, wherein said power law function is defined as:

$r_k = (K/(N+1))^{(1/S)}$, wherein $S \in (1, \infty)$.

ABSTRACT

The delivery ratio of r (which is a fraction between 0 and 1) partitions a stream of documents into a section of top scoring r-fraction of documents and the remainder. This way a set of successively bigger delivery ratios, $r_1, r_2, r_3, \ldots$ sections the stream into tiers. Any given document is assigned to a tier according to how many delivery ratio thresholds it matched or surpassed and how many it failed to reach. This creates a scoring structure which reflects the specificity of the document with respect to a profile in terms of density of relevant documents in the stream. In other words, a document in the k^{th} tier is such that it failed to be classified in the top r_k ratio of the stream (thus r_k fraction of the stream is more relevant to the given profile than the document under consideration). At the same time this document was classified as being in the top r_{k-1} part of the stream. Thus this mechanism defines a score (let's call it σ) for a document depending on how it compares to other documents in the stream when scored against a given profile.

Figure 1

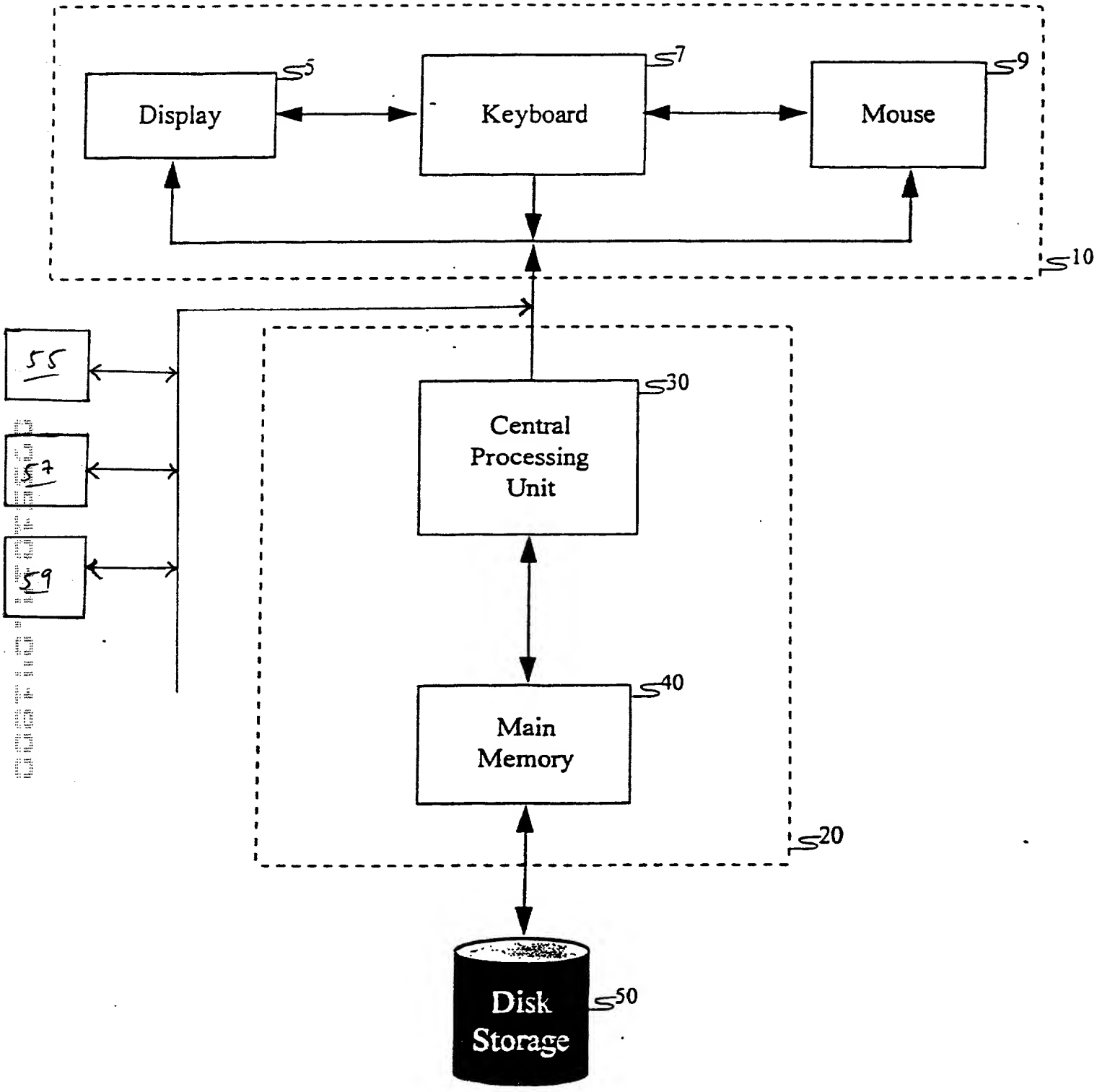

Display
5
Keyboard
7
Mouse
9
10
55
57
59
Central Processing Unit
30
Main Memory
40
20
Disk Storage
50

Figure 2

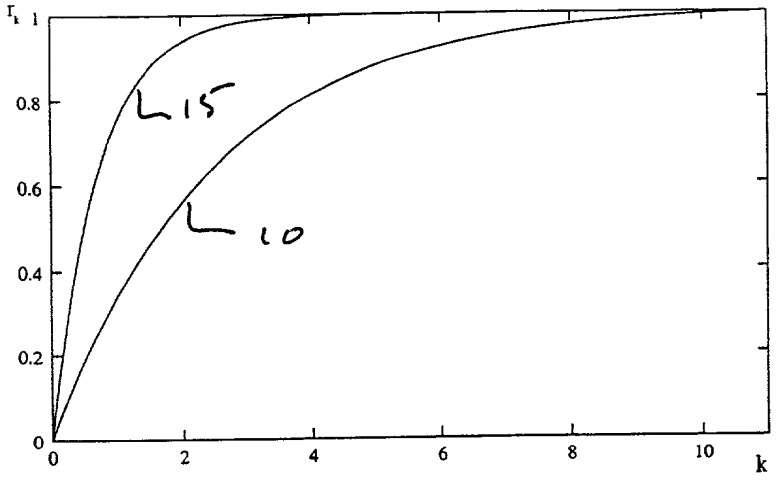
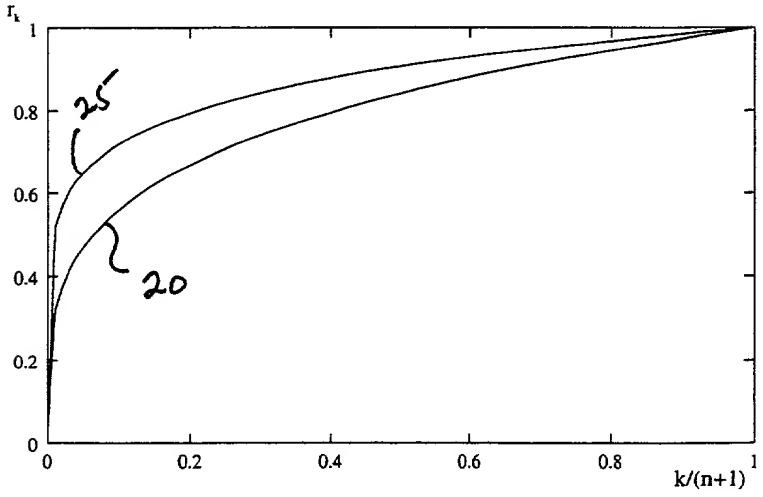


Figure 3



COMBINED DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name.

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

A METHOD AND APPARATUS FOR COMPARING SCORES IN A VECTOR SPACE RETRIEVAL PROCESS

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material of the examination of this application in accordance with Title 37, Code of Federal Regulations, Section 1.56(a).

I hereby claim foreign priority benefits under Title 35, United States Code, Section 119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed.

Prior Foreign Application(s)		Priority Claimed	
Number _____	Country _____	Yes ____	No ____
	Day/Month/Year Filed _____		
Number _____	Country _____	Yes ____	No ____
	Day/Month/Year Filed _____		
Number _____	Country _____	Yes ____	No ____
	Day/Month/Year Filed _____		

I hereby claim the benefit under Title 35, United States Code, Section 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, Section 112, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, Section 1.56(a) which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

Application Serial No.: _____ Filing Date:_____
Status: _____

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

I hereby appoint Blaney Harper, Registration No. 33,897, as our attorney, with full power of substitution and appointment of associate attorneys, to prosecute this application and to transact all business in the Patent and Trademark Office connected therewith.

Send all future correspondence to:

Blaney Harper, Esq.
Jones, Day, Reavis & Pogue
51 Louisiana Avenue, N.W.
Washington, DC 20001-2113

* * *

Full name of sole or first inventor:

NORBERT ROMA

Inventor's signature _____

Date: , 19

Residence: 115 Conover Road
Pittsburgh, Pennsylvania 15208
Citizenship: Poland
Post Office Address: Same as above.

Full name of second joint inventor:

Inventor's signature _____

Date: , 19

Residence:

Citizenship: United States of America
Post Office Address: Same as above.

Full name of third joint inventor:

Inventor's signature _____

Date: , 19

Residence:

Citizenship: United States of America
Post Office Address: Same as above.